Policy Change Index: A Simulated Example

December 3, 2018; Version 1.0

Julian TszKin Chan and Weifeng Zhong*

Chan Bio: Julian TszKin Chan is a senior economist at Bates White Economic Consulting. His research focuses on econometrics, machine learning, and their applications to network and text data. Julian is a core maintainer of the open-source PCI models.

Zhong Bio: Weifeng Zhong is a research fellow at the American Enterprise Institute and an expert at the Open Research Group, Inc. His research focuses on political economy. His recent work has been on the applications of text analysis and machine learning to Chinese economic issues. Weifeng is a core maintainer of the open-source PCI models.

Abstract: *In a recent research paper, we use the text of the People's Daily to develop the Policy Change Index (PCI) for China, a machine-learning-based predictor of China's policy changes. In this Quantitative Note, we provide an example in which both the text data and the change in priorities are simulated. We apply the PCI design to show that the algorithm can correctly identify the simulated change. This example allows the reader to understand and experiment with the PCI design without relying on the People's Daily text.*

China's industrialization process has long been a product of government direction. Yet, the opaque political system makes it prohibitively challenging to measure the Chinese government's policy priorities, not to mention predicting their changes.

For the first time in the literature, we develop a leading indicator of China's policy priorities, the Policy Change Index (PCI) for China, in Chan and Zhong (2018c).[1] We construct the PCI by building a machine learning algorithm to "read" the *People's Daily* — China's official newspaper — and detect changes in its front-page content. Due to the highly official status of the *People's Daily* (see Wu, 1994), the changes in the way it prioritizes articles provide valuable clues as to how policies *will* change in the future. The PCI for China covers the period from 1951 to as long as the *People's Daily* stays in print. It not only has successfully "predicted" many of China's major policy changes in the past but is also able to make policy predictions in the future.

While we have released the source code[2] of the PCI model, some researchers may find it difficult to replicate our findings or build their research on our method due to data availability issues. In this *Quantitative Note*, we provide an example with simulated data and demonstrate the PCI design without requiring the availability of the *People's Daily* text. The interested reader can also experiment with different specifications of the simulated example. The source code of the simulated example can be found in the project's repository.[3]

1. Design of the PCI for China

The design of the PCI algorithm, described in detail in Chan and Zhong (2018c), mimics the mind of a China observer who avidly reads the *People's Daily* and stays vigilant about its content. If the observer had read and thought through all the articles published in recent times, they would have formed a fairly rigorous paradigm about what articles are important enough to "deserve" the front-page status and what articles are not. This thought experiment essentially reverse-engineers the *People's Daily* editor's mind.

But if the observer then woke up to a surprising paper one day — that is, their educated guess about the new paper turned out either particularly well or exceptionally poorly — that would be a signal of change from the observer's perspective. One may well deem a small surprise as noise, but a strong signal would convince the observer that their existing understanding of the front-page content is no longer valid and that a fundamental change must have occurred to the priorities of the *People's Daily*. In other words, an unusual error rate in this

[*] *Quantitative Notes* is published by Open Research Group, Inc., a public benefit corporation. For more *QNs* or to arrange meetings with an expert, please contact experts@openrg.com. The authors did not receive funding from any source for the production of this *Quantitative Note*. The views expressed here are solely those of the authors and do not represent the views of the American Enterprise Institute, Bates White Economic Consulting, or their other employees.

[1] Also see our project website https://policychangeindex.com.

[2] See our repository https://github.com/open-source-economics/PCI.

[3] See this GitHub page in our repository.

context indicates a structural change in the data-generating process.

The construction of the PCI takes two steps. First, we train the machine with a set of *People's Daily* articles from a certain time window as well as whether they appear on the front page. The machine uses these training examples to teach itself how to perform front-page classification. Second, we test the algorithm's performance on new data that are unseen by the machine and that may or may not be structurally similar to the training data. If the performance on new data does drastically differ from the performance on the training data, that would mean the new data are indeed driven by different priorities.

As researchers, we do not need to understand the mind of the *People's Daily* editor. We do not need to understand the mind of the China observer either. Just the magnitude of surprise the observer gets is sufficient for us to conclude a structural change underlying the *People's Daily*'s priorities.

2. A Simulated Example

In this section, We illustrate the construction of the PCI using a simulated data set. The example abstracts from the complex Chinese context — its language, economy, politics, etc. — and allows the reader to grasp the mechanism of detecting changes in the data-generating process.

2.1 Data

We simulate articles using numerical digits. Imagine that an article is simply a sequence of ten single digits (from "0" to "9"), such as $(4,1,8,7,9,2,5,...)$. Each digit is analogous to a Chinese character; after all, the difference does not matter to the machine. Each article is simulated by drawing ten digits randomly and independently. We simulate 1,000 articles in each period, for ten periods.

Next, we simulate the editor's behavior, which is supposedly unknown to the machine. We assume that, for some reason, the editor assigns page numbers to the articles in the following manner. Throughout the first five periods, for each article, the editor looks for the first appearance of digit "9." Once the first "9" is found, the article is assigned to the front page if the next digit is larger than "5." The article is assigned to other pages otherwise. The example mentioned above, article $(4,1,8,7,9,2,5,...)$, would not be assigned to the front page because the digit following the first "9" is "2," which is lower than the threshold "5."

One can interpret this behavior as the editor looking for the name of the Chinese president (digit "9") and then examining whether the activity the president is involved in (the next digit) is high-profile enough for the front page.

Suppose, starting from period 6, the editor changes their behavior and starts looking for the first appearance of digit "8," instead of "9." The rule regarding the next digit remains the same. This might mean, for example, the cue the editor pays attention to is no longer the president (digit "9"), but the premier (digit "8").

The problem posted to the algorithm we will build next is whether it can correctly identify that period 6 is when a change in the data-generating process occurs.

2.2 Machine Learning Model

For each period, we train a separate model to learn — or, reverse-engineer — the editor's decision rule.

For the 1,000 articles we simulated for each period, we randomly split them into the set of training data and the set of testing data. The former are used to train a model, while the latter, as well as the trained algorithm, will be used to construct the PCI.

The training data in period t are used to fit the following function:

$$f_t : X \to Y, \tag{1}$$

where $X \in \{0,1,...,9\}^{10}$ is the set of all possible articles and $Y \in \{0,1\}$ is the front-page indicator. Therefore, f_t is the editor's decision rule in period t.

The neural network model we use to fit f_t is a simplified version of the one in Chan and Zhong (2018c), where more details can be found. It consists of an embedding layer, which extracts features behind digits, and a gated recurrent unit (see Cho et al., 2014), a type of recurrent neural networks that processes the digits sequentially — just like the way humans read text.

2.3 Construction of the PCI

To construct the PCI, we first apply the model for each period on two sets of data: the testing data in the same period where the model was trained and the data in the next period, which we call the "one-step-ahead" data. We then compare the performance of the model in these two sets of data.

To asses how well the model is fit in the first place, we apply the model to the same-period testing data. The rationale for using the testing data, instead of the training data, to assess the fitness is to avoid the infamous over-fitting problem. That is, if a modeler maximized the fitness of the model based on the training data, the optimized model would be overly influenced by the idiosyncratic features contained in the training data but *not* in other data, such as the testing data. Over-fitting the model on the training data would then worsen the model's out-of-sample predictive power, which is the goal of training the model in the first place.

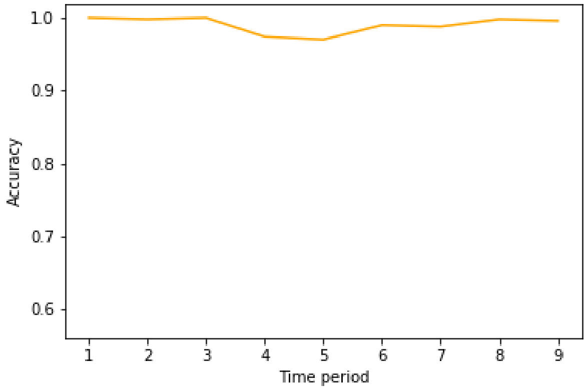

Figure 1. Testing Accuracy

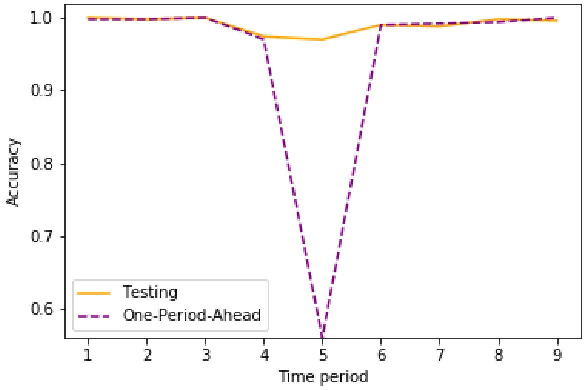

Figure 2. Testing Accuracy and One-Step-Ahead Accuracy

Figure 1 plots the Testing Accuracy — the accuracy rate[4] calculated using the same-period testing data.[5] Because the editor's decision rule we simulated is relatively straightforward, it is no surprise that the accuracy rate is nearly perfect.

Next, we apply the model for each period on the respective "one-step-ahead" data. If the data-generating process, which is unknown to the modeler, does not change from one period to another, the model's performance should be fairly close to the Testing Accuracy. But if they drastically differ from each other, that would imply the data-generating process has fundamentally changed.

Figure 2 shows both the Testing Accuracy and the One-Step-Ahead Accuracy — the accuracy rate calculated using the "one-step-ahead" data. The One-Step-Ahead Accuracy is very close to the Testing Accuracy, *except for period 5*, when it drops drastically. This implies that the underlying decision rule that determines whether an article appears on the front page changes in the next period — period 6 — which coincides exactly with how we simulated the data in the first place.

Finally, the PCI is constructed by taking the difference in absolute value between the Testing Accuracy and the One-Step-Ahead Accuracy, which is the difference between the two curves in Figure 2. Figure 3 plots the PCI for our simulated data. When the index's value is high, it indicates a structural change in the editor's decision rule, which was unknown to the algorithm by design but has been inferred correctly now.

2.4 Variations of the Simulated Example

The interested reader can experiment with different variations of the simulated example given here. For example, instead of a change in the cue digit (i.e., from "9" to "8"), one can simulate a change in the threshold following the cue (e.g., from "5" to "6"). Alternatively, the editorial criterion for appearing on the front page may depend on multiple digits before or after the cue. We relegate to the interested reader to verify that the PCI can pick up structural changes in these circumstances as well.

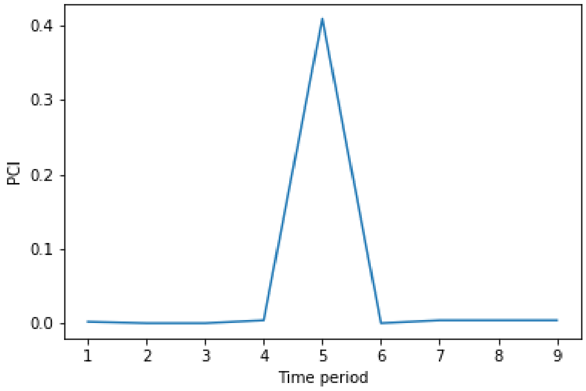

Figure 3. Policy Change Index (simulated)

3. Advantages and Potential Applications

In this section, we discuss the advantages of the PCI design and explore potential applications to which it can be applied.

In our original application, the label — the dependent variable of the function f_t — is an indicator of whether each article appears on the front page. This label is not only substantively important in the subject matter but also easily available in the data set as a metadata field. Our research design exploits this fact and reduces the cost of constructing the training data set — to virtually zero. We discuss this property of our method in more depth in Chan and Zhong (2018b).

[4] The accuracy rate is defined as the fraction of correct classifications in all (testing) data.

[5] There are no "one-step-ahead" data to which the model for period 10 can be applied. We thus omit the data point for period 10 in the figure.

Another advantage of the PCI design is that it minimizes the bias a researcher may have regarding the subject matter, such as the Chinese economy and reforms. Instead of studying the newspaper's priorities directly, which would necessarily require a researcher to characterize the priorities and likely inject their views and biases in the meantime, our method bypasses that and investigates *the structural changes of* the priorities instead. We discuss in more detail this "twist" in our method in Chan and Zhong (2018a).

The PCI design has a variety of potential applications, the most obvious of which is the construction of PCIs for other (ex-)Communist regimes. Just like China, these regimes — such as the Soviet Union, East Germany, North Korea, Cuba, and Vietnam — publish(ed) their official newspapers and utilize(d) them to achieve the government's policy goals. We are hoping to soon demonstrate this application with our work in progress, the PCI for Cuba.

Page numbers of newspaper articles are by no means the only example of easily available labels that are of subject-matter importance. Names of people are another example. A potential application along this line would be to study legislators' public statements and take their names as the easily available labels. A design similar to the PCI would then be able to detect whether a legislator's statements are mistaken by the algorithm as coming from another legislator, suggesting a structural change in the former's stance. Just like the Chinese government, which changes its propaganda before changing its policies, legislators may as well change their public statements before changing their votes. This approach, therefore, would potentially allow one to create a predictor of voting behavior.

In our research paper (Chan and Zhong, 2018c), we explore more potential applications of the PCI design, to which the interested reader can turn.

References

Chan, Julian TszKin and Weifeng Zhong, "Machine Learning with a Twist: Detecting Structural Differences in Complex Data," *MLconf*, 2018. Available at: https://mlconf.com/machine-learning-with-a-twist-detecting-structural-differences-in-complex-data/.

_ **and** _ , "Machine Learning with a Twist: How Trivial Labels Can Be Used to Predict Policy Changes," *Dataconomy*, 2018. Available at: http://dataconomy.com/2018/11/machine-learning-with-a-twist-how-trivial-labels-can-be-used-to-predict-policy-changes/.

_ **and** _ , "Reading China: Predicting Policy Change with Machine Learning," AEI Economics Working Paper 2018-11, American Enterprise Institute, Washington, DC 2018.

Cho, Kyunghyun, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing" 2014.

Wu, Guoguang, "Command Communication: The Politics of Editorial Formulation in the People's Daily," *The China Quarterly*, 1994, *137*, 194–211.